پروژه پایانی پنجمین مدرسه مقدماتی علم داده و هوش مصنوعی

دو مجموعه داده از آگهی های فروش ماشین های کارکرده، جمع آوری شده از یک پلتفرم داخلی در اختیار شما قرار داده شده است. مجموعه داده آموزشی (Train) دارای ستون قیمت و حاوی مقادیر قیمت اعلام شده برای هر ماشین می باشد. می خواهیم یک مدل یادگیری ماشین، با استفاده از روش های یادگیری با ناظر (Supervised Learning) بسازیم، که با دریافت اطلاعات ماشین (Features)، تخمینی قابل قبول برای قیمت آن را محاسبه کند.

مجموعه داده دوم نیز به عنوان مجموعه تست (Test Dataset) در اختیار شما قرار گرفته است؛ این مجموعه فاقد اطلاعات قیمت می باشد و فقط حاوی اطلاعات هر آگهی (Features) می باشد. مدل های ساخته شده توسط شما با استفاده از این مجموعه تست، ارزیابی خواهند شد. قیمت هر ماشین در مجموعه داده تست باید توسط مدل آموزش دیده شما پیش بینی شود.

ویژگی های موجود از هر ماشین به صورت زیر است:

- title: برند و مدل ماشین
- year: سال ساخت ماشین
- mileage: کیلومتر کارکرد
- transmission: نوع گیربکس
- fuel: میزان مصرف سوخت
- body_color: رنگ بدنه
- inside_color: رنگ داخل
- body_status: وضعیت بدنه
- description: متن توضیحات آگهی
- body_type: نوع بدنه
- volume: حجم موتور
- engine: مشخصات موتور
- acceleration: زمان صفر تا صد
- price: قیمت اعلام شده

برخی نکات مهم برای حل بهتر مسئله:

- از انجام پیش پردازش و EDA کامل و خالقانه بر روی دادگان train غافل نشوید.
- توجه شود که به خاطر ماهیت آگهی های اینترنتی، امکان وجود نمونه هایی با قیمت های بسیار نادرست (noisy)، در داده train وجود دارد. سعی کنید برای حذف اینگونه noise ها راهکاری بیاندیشید.
- در متن description آگهی، معمولا اطلاعات ارزشمندی در مورد ماشین ذکر شده است، استفاده از این اطلاعات به عنوان ورودی مدل یادگیری ماشین ممکن است در تخمین بهتر قیمت اثرگذار باشد. البته چگونگی استفاده از متن description به عهده شماست.
- روش های مهندسی ویژگی (Feature Engineering)، مثل تبدیل یک ویژگی به چند ویژگی، یا محاسبه یک ویژگی از یک یا چند ویژگی دیگر، ممکن است کمک کننده باشند.

نحوه و معیار های ارزیابی

در این پروژه، خروجی نهایی کار شما، یک نسخه از فایل دادگان تست (Test) خواهد بود، که ستون price آن را با مقادیری که مدل یادگیری ماشین شما تخمین زده است، پر کرده اید. ما با محاسبه متریک های ارزیابی ای که در ادامه درباره آن ها صحبت خواهیم کرد، نمره ای برای شما محاسبه می کنیم. در ارزیابی مسائل regression، متریک های گوناگونی از جمله MAE و R2 مورد استفاده قرار می گیرند. در این مسئله، ما به خاطر ماهیت مسئله قیمت

گذاری آگهی، به مقادیر Absolute Error کاری نداریم؛ آنچه برای ما اهمیت دارد، میزان خطای نسبی است. خروجی شما با دو متریک متفاوت ارزیابی خواهد شد که در ادامه هر کدام را معرفی می کنیم:

۱. میانه درصد خطا (Median Absolute Percentage Error)

میانه درصد خطا به صورت زیر محاسبه می شود:

$$Median\left(\frac{|y_{true} - y_{pred}|}{y_{true}}\right) * 100$$

این متریک نشان می دهد که به صورت میانگین، مقادیر پیش بینی شده توسط مدل چند درصد از واقعیت فاصله دارند. (البته در اینجا به جای میانگین از میانه استفاده می کنیم) توجه شود که این متریک خطا را نشان داده و هر چه به صفر نزدیک تر باشد به معنی بهتر بودن پیش بینی های مدل است.

۲. R2

این متریک یک اندازه گیری آماری از انطباق پیش بینی های مدل رگرشن بر واقعیت را ارائه می کند. مقدار ۱ نشان دهنده یک مدل ایده آل است.

امتیاز نهایی

امتیاز نهایی شما میانگینی از دو متریک معرفی شده است که به صورت زیر محاسبه می شود:

$$\frac{(1 - MAPE) + R2}{2} * 100$$

توجه شود که محاسبه متریک های ذکر شده و امتیاز نهایی بصورت خودکار در بستر پلتفرم سکو و توسط هسته ارزیابی آن برای پیش بینی های مدل شما محاسبه خواهد شد و شما ملزم به پیاده سازی این متریک ها نیستید.

نحوه ارسال پاسخ

پس از ساخت یک مدل یادگیری ماشین، تمام رکورد های موجود در مجموعه تست را، بدون هیچ گونه تغییر در ترتیب، یا حذف آن ها، به مدل خود داده و مقدار قیمت را برای آن ها محاسبه کنید. سپس قیمت های حاصله را با همان ترتیب در یک فایل csv و در ستونی با نام price ذخیره کرده و در قسمت ارسال پاسخ در پلتفرم سکو آپلود کنید. سکو بلافاصله پس از آپلود پاسخ، آن را ارزیابی کرده و امتیاز پاسخ ارسالی را ثبت می کند.